Genomics xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Computational dissection of *Arabidopsis* smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites

Xiangfeng Wang [a,b,*], John D. Laurie [a], Tao Liu [b], Jacqueline Wentz [c], X. Shirley Liu [b,**]

[a] School of Plant Sciences, University of Arizona, 1140 E. South Campus Drive Tucson, AZ 85721, USA
[b] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02115, USA
[c] Department of Bioengineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 12 October 2010
Accepted 27 January 2011
Available online xxxx

Keywords:
High-throughput sequencing
Small RNAs
Principal component analysis
TSS-associated RNAs

ABSTRACT

The profiling of small RNAs by high-throughput sequencing (smRNA-Seq) has revealed the complexity of the RNA world. Here, we describe a computational scheme for dissecting the plant smRNAome by integrating smRNA-Seq datasets in *Arabidopsis thaliana*. Our analytical approach first defines *ab initio* the genomic loci that produce smRNAs as basic units, then utilizes principal component analysis (PCA) to predict novel miRNAs. Secondary structure prediction of candidates' putative precursors discovered a group of long hairpin double-stranded RNAs (lh-dsRNAs) formed by inverted duplications of decayed coding genes. These gene remnants produce miRNA-like small RNAs which are predominantly 21- and 22-nt long, dependent of DCL1 but independent of RDR2 and DCL2/3/4, and associated with AGO1. Additionally, we found two classes of transcription start site associated (TSSa) RNAs located at sense (+) and antisense (−) approximately 100–200 bp downstream of TSSs, but are differentially incorporated into AGO1 and AGO4, respectively.

Published by Elsevier Inc.

1. Introduction

Plant genomes produce a variety of small RNA (smRNA) families to mediate either post-transcriptional or transcriptional gene silencing (PTGS or TGS). In *Arabidopsis*, three known classes of small RNAs functioning in PTGS comprise microRNAs (miRNAs), *trans*-acting siRNAs (tasiRNAs) and natural antisense transcript-derived siRNAs (natsiRNAs) that guide the cleavage of mRNAs [1–4]. The fourth class of endogenous siRNAs acting in TGS arises from the transposable elements (TEs) to mediate the epigenetic silencing of cognate TEs [5–8]. Those small RNAs are recently uniformly defined as *cis*-acting siRNAs (casiRNAs) [9]. Functional categorization of those small RNAs is based on their distinct mechanisms of biogenesis by a combination of different members of RNAi components encoded in *Arabidopsis* genome, which include four Dicer-like endonuleases (DCL1–4) [10], Pol II and other two plant-specific DNA-dependent RNA polymerases, Pol IV and Pol V [11,12], six RNA-dependent RNA polymerases (RDR1–6) and ten Argonautes (AGO1–10) [13].

Transcription of a miRNA gene (*MIR*) is dependent on Pol II. The primary transcript of a *MIR* gene is a long single-stranded RNA called pri-miRNA that contains an imperfect inverted repeat and is further cleaved into precursor miRNA (pre-miRNA) with a stem–loop structure. In plants, the two steps of processing from the pri-miRNAs to pre-miRNAs, and to mature miRNA duplexes are catalyzed by DCL1 [14]. While the guide strands of the miRNA duplexes are incorporated into AGO1 of the RNA-induced silencing complex (RISC), the passenger strands called miRNA star (miRNA*) are mostly degraded. Plant miRNAs are typically 21-nt long, preferentially started with a uracil at 5′ end. Unlike the animal miRNAs that target mRNA's 3′ UTR by the "seed regions (the 2nd to 8th nucleotide from a miRNA's 5′ end)", plant miRNAs are usually complementary to their targets' coding regions with near-perfect match to induce the cleavage [14].

In plants, tasiRNAs are discovered to have the similar function with miRNAs to regulate the gene silencing at posttranscriptional level, but in a manner of imperfect matching with their targets [15]. The genomic loci encoding tasiRNAs are known as *TAS* genes transcribed by Pol II, and the mature tasiRNA products are uniformly 21-nt long started with a U at 5′ ends. The third class of siRNAs in PTGS is natsiRNA whose long dsRNA precursors are formed by the hybridization of overlapping sense and antisense RNA transcripts caused by convergently transcribed genes or TEs [16].

In plants, casiRNAs are the most predominant class of small RNAs and are prevailingly produced from transposable elements, heterochromatic regions or other repetitive sequences. Therefore, casiRNAs are previously called TE-derived siRNAs, heterochromatic siRNAs (hcRNAs) or repeat-associated siRNAs (rasiRNAs) [4,7]. The functional role of casiRNAs is to direct the DNA methylation on the genomic loci where they originate from and silence the residing TEs in *cis* [17]. It also has been indicated that casiRNA pathways might influence the transcription of the neighboring protein-coding genes as they can

* Correspondence to: X. Wang, School of Plant Sciences, University of Arizona, 1140 E. South Campus Drive Tucson, AZ 85721, USA.
** Corresponding author.
E-mail addresses: xwang1@cals.arizona.edu (X. Wang), xsliu@jimmy.harvard.edu (X.S. Liu).

0888-7543/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.ygeno.2011.01.006

modify the epigenetic states of upstream sequences [18,19]. The casiRNAs possess two signatures, 24-nt long and preferential A at 5′ end, which can be recognized by AGO4, a component of RNA-directed DNA methylation (RdDM) complex.

The high-throughput profiling of small RNAs by sequencing (smRNA-Seq) has revealed the complexity of the RNA population. Those exponentially accumulating smRNA-Seq datasets have created urgent challenges for quantitative interpretation of the results and *in silico* identification of new smRNA classes and pathways. In addition to those known miRNAs, tasiRNAs, natsiRNAs and casiRNAs, many functionally uncharacterized small RNAs have been observed to arise from structured genomic sites such as long inverted repeats, short hairpin repeats, and convergent genes, whose biogenesis pathways may differ from canonical mechanisms. Recently, several software packages and pipelines have been developed to cope with the large-scale analysis of smRNA-Seq datasets mainly aiming at two purposes: first, to process raw smRNA-Seq data and annotate the small RNAs in the genome; second, to build the expression profiles of known miRNAs and discover the new miRNAs [20–25].

The first way to identify new miRNAs from smRNA-Seq data is based on cross-species comparison, which is to directly align the reads with known miRNAs in other species such as adopted by miRExpress and DSAP [20,21]. The other way is to find new miRNAs according to the miRNA biogenesis pattern which is the features of how miRNA mature products are processed from pre-miRNA hairpin precursors. The original algorithm was developed by Friedländer et al. and was implemented as a software package called miRDeep [22]. miRDeep first extracts putative miRNA precursors with uniquely mapped smRNA reads and then rules out those overlapped with rRNA, snoRNA, tRNA loci etc., as well as those that cannot fold into canonical hairpin structures [22]. Next, miRDeep uses Bayes' theorem to calculate the probability of a potential miRNA precursor by comparing with background hairpins [22]. The algorithm of miRDeep was also integrated by other smRNA-Seq analysis tools to identify the new miRNAs such as deepBase and mirTools [23,24]. Another *de novo* miRNA prediction tool, miRanalyer utilizes machine learning approach to score the new miRNAs based on a variety of features such as read counts, stem and loop lengths, and folding energy etc. [25].

As miRNA is the predominant type of small RNAs in animals, most available smRNA-Seq tools focus on miRNA analysis. Although the basic concepts of miRNA prediction from smRNA-Seq are essentially the same for animals and plants, notable differences still exist. For example, while the animal miRNA precursors have more canonical hairpin structures with relatively fixed size of stem and loop regions, plants pre-miRNAs sometimes have longer hairpin stem regions and even multiple branches. Additionally, plant genomes contain a great number of inverted repeats formed by transposable elements that produce miRNA-like siRNAs, which are usually the source of false positive results from *de novo* miRNA prediction. Furthermore, as the majority of plant small RNAs are various types of siRNAs, a more comprehensive pipeline needs to be developed to annotate existing siRNAs and discover the new species. By integrating six smRNA-Seq datasets in different developmental stages and RNAi pathway mutations [26–30] (Supplementary Table 1 and Supplementary Fig. 1), we developed an analytical framework to dissect the *Arabidopsis* smRNAome and computationally discover previously uncharacterized miRNAs and other smRNA classes.

2. Materials and methods

2.1. Define smRNA-deriving loci as primary transcription units (Pri-TU)

We obtained the four libraries of processed Argonaute-associated (AGO1, AGO2, AGO4 and AGO5) smRNA-Seq dataset from Dr. Yijun Qi's group, in which the 5′ and 3′ adaptor sequences had been trimmed off from both ends of the sequencing reads. This dataset contains totally 2,840,770 high-quality reads that represent 599,449 unique small RNA sequences.

To determine the genomic locations of small RNA reads, we employed Bowtie [31] to map the ~600,000 unique smRNA sequences to *Arabidopsis* reference genome TAIR8 (http://www.arabidopsis.org/), and kept all locations that a read was perfectly aligned to. By bowtie, 599,397 of them were mapped to 2,654,309 locations without any mismatch. Thus, each unique small RNA sequence has two layers of information: (1) the *repetitiveness*, the number of the locations it was mapped to the genome without any mismatches, and (2) the *abundance*, the number of the reads for a unique small RNA being sequenced.

We developed a tool to *de novo* scan the genomic mapping result of smRNA-Seq reads to define the primary transcription units (Pri-TUs) that give rise to small RNAs. As Fig. S2 shows, for a putative Pri-TU, it was composed of a set of small RNAs that are overlapped or next to each other with a small gap (Supplementary Fig. 2). The initial *de novo* scanning identified 108,350 Pri-TUs with maximum 50 bp gap allowed, and at least 2 reads per Pri-TUs. Since most of the Pri-TUs containing very few reads might be resulted from the wrong mapping or background noise, we only used 23,516 Pri-TUs containing more than 20 reads for the further statistics.

During the identification of Pri-TUs, we also collected following information for each Pri-TU: (1) *SeqFreq* (sequencing frequency), which is the sum of the reads that a small RNA were being sequenced, to represent the expression abundance of a small RNA; (2) *RepFreq* (repetitive frequency), which is the sum of all the locations for a small RNA whose sequence was mapped in the genome, to represent the repetitiveness of a small RNA; (3) *UniqFreq* (unique frequency), which is the sum of the number of unique smRNA sequences within a Pri-TU, to represent the excision mode; (4) *AvgSeq* is the ratio of *SeqFreq*/*RepFreq*, which is the adjusted value of small RNA abundance by repetitive frequency; (5) *size* and (6) *5′ terminal-nt* is the most prevalent length and the type of 5′ terminal nucleotides of the small RNAs inside a Pri-TU, respectively. After the Pri-TUs were identified, we also calculated the following features including the frequency of the cutting sites of di-nucelotide where small RNAs were processed from Pri-TU, the proportions of 5′A, 5′G, 5′C and 5′U, the strand-bias that small RNA derived from plus and minus strand within a Pri-TU (Supplementary Table S2).

2.2. Computational selection of candidate Pri-TUs for new miRNA prediction by principal component analysis (PCA)

Computational selection of candidate Pri-TUs was based on the facts that miRNAs tend to be sequenced more (higher *SeqFreq*), but more accurately excised from pre-miRNA hairpins, and uniquely mapped in the genome (lower *RepFreq* and *UniqFreq*). We employed principal component analysis (PCA) on *SeqFreq*, *RepFreq* and *UniqFreq* to discriminate the Pri-TUs of producing miRNAs from the ones producing siRNAs [32]. The nature of PCA algorithm is to identify the direction (first principal component, PC1) with the largest variation, and the direction of the second and third principal components (PC2, PC3) uncorrelated to PC1. The three PCs were standardized to be centered at zero, and we used $PC1>0$, $PC2<0$ and $PC3<0$ to classify the miRNA-deriving Pri-TUs and siRNA-deriving Pri-TUs. After removing Pri-TUs associated with known miRNA genes, the rest candidate Pri-TUs will be used for further new miRNAs prediction. We next searched the candidate Pri-TU sequences against *Arabidopsis* TAIR8 annotation to further exclude the false positive candidates which were tasiRNAs, snRNAs, snoRNAs, tRNAs and rRNAs etc. whose secondary structure may contain hairpins. The second round screening narrowed the candidates down to those were absolutely located in the intergenic regions based on TAIR8's annotation. We then extracted the precursor sequences by extending at 35 bp on both end of a Pri-TUs to predict their secondary structures by RNAfold.

Please cite this article as: X. Wang, et al., Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites, Genomics (2011), doi:10.1016/j.ygeno.2011.01.006

RNAfold calculates the Minimum Free Energy (MFE) for each candidate Pri-TU, and those Pri-TUs whose MFEs significantly lower than background energy were considered as new miRNA genes.

3. Results and discussion

3.1. Computational classification of plant small RNAs from RNA-Seq data

Classification of plant miRNAs and various endogenous siRNAs is based on their distinct biogenesis pathways and regulated targets [33]. Each class has a distinct pattern of DICER excision from their double-stranded (ds) RNA precursors (Fig. 1A). Armed with this knowledge, we developed a computational pipeline to process the smRNA-Seq reads mapping results of Bowtie [31] and to define a cluster of overlapped or slightly gapped smRNA-Seq reads as a primary transcription unit (Pri-TU) encoding an initial RNA transcript (Fig. 1B). Using Pri-TUs as basic units allows us not only to perform normalization and comparisons between libraries but also to classify different types of smRNAs based on their genomic locations, sequence and structural characteristics (Supplementary Fig. 2). When applied to a recently published deep-sequencing dataset of the small RNAs extracted from purified AGO1, AGO2, AGO4 and AGO5 complexes (RIP-Seq) [26] in the *Arabidopsis* genome (Supplementary Fig. 3), the pipeline identified a total of 108,350 Pri-TUs. 23,516 Pri-TUs with over 20 reads were selected for subsequent statistical analysis (Supplementary Table 2).

We studied the sequence characteristics of the smRNAs in these Pri-TUs. The average length of the 23,516 Pri-TUs is 403 bp, whose distribution was shown in Supplementary Fig. 4. The majority of the Pri-TUs (83.2%) preferentially produce 24-nt smRNAs, which are mostly *cis*-acting siRNAs (casiRNAs) functioning in RNA-directed DNA methylation (RdDM) mechanism to suppress the transposable elements (TEs) (Fig. 2A). Interestingly, analysis of AGO RIP-Seq data indicates that AGO4 can preferentially associate with not only the longer smRNAs of 23–27 nt, but also the shorter ones of 19 and 20 nt (Fig. 2B). As the 5′ terminal nucleotide (5′nt) of a smRNA dictates its preferred AGO association [26], we examined the type of preferential 5′nt in smRNAs of different sizes. While 5′A is prevalent among 23- and 24-nt smRNAs, the shorter (19- and 20-nt) and longer (25- and 26-nt) smRNAs tend to initiate with 5′G (Fig. 2C). The difference between casiRNAs and miRNAs is that the former class is yielded from TEs by semi-random excisions of long dsRNAs, but the latter is from unique *MIR* genes by precise excision of the stem region of a small hairpin RNA [34]. We therefore investigated the correlations of smRNA sizes and three frequencies of the smRNA reads within a given Pri-TU: number of times being sequenced (*SeqFreq*), total number of mapped locations (*RepFreq*), and the number of the unique smRNA sequences (*UniqFreq*) (Fig. 2D). Surprisingly, 23-nt class Pri-TUs demonstrated the highest *RepFreq*, most of which are actually composed of ~50-nt Poly-A or Poly-T (Supplementary Table 3). While the reads from 23- and 24-nt class Pri-TUs typically have hundreds to thousands of mappable locations, 21- and 22-nt smRNAs are sequenced with the highest frequencies, because most of them are from known miRNA genes (Fig. 2D). At last, examination of the most frequent di-nucleotide cutting sites showed that A|A and U|U has the highest chance to be diced, as well as A|U and U|A in the second preference (Fig. 2E and Supplementary Fig. 5). We also developed extension modules to align the Pri-TUs with the annotated genomic compartments in TAIR8, such as TEs, housekeeping RNA genes, protein-coding genes (Supplementary Figs. 6 and 7).

3.2. Computational selection of new miRNA candidates by PCA analysis

We then used the known miRNA genes as a training set to model the characteristics of miRNAs and siRNAs in expression abundance (*SeqFreq*), mapping uniqueness (*RepFreq*) and excision accuracy (*UniqFreq*). Interestingly, the known miRNA-deriving Pri-TUs tend to have higher *SeqFreq*, but lower *RepFreq* and *UniqFreq* (Fig. 3A and B). We therefore conducted principal component analysis (PCA) [32] on 9254 Pri-TUs with over 100 reads to discriminate Pri-TUs harboring miRNAs from those harboring siRNAs, and select the

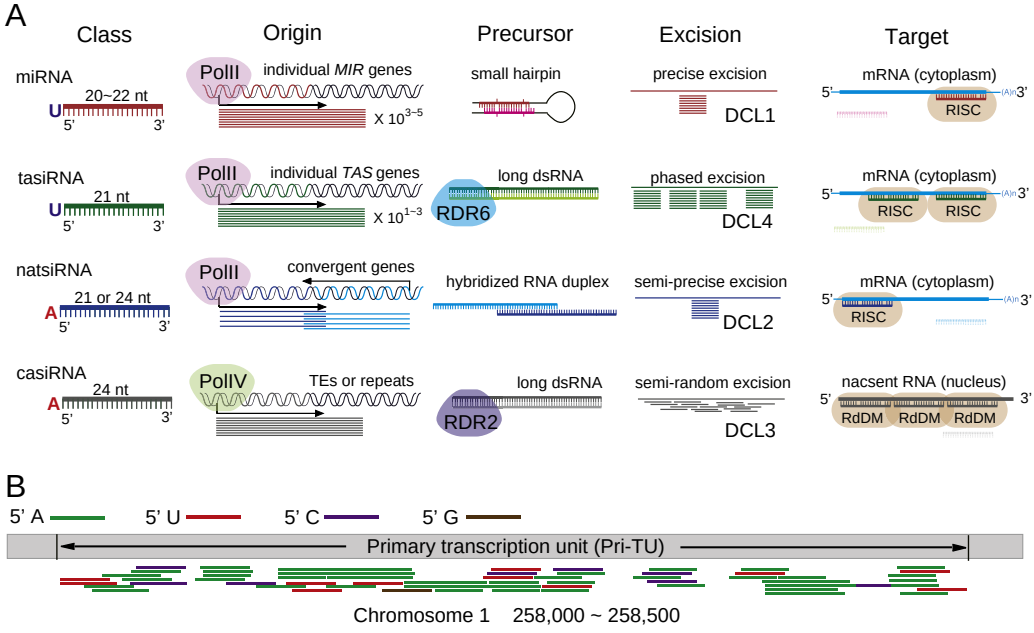

Fig. 1. Define primary transcription units (Pri-TUs) that produce small RNAs. (A) Classification of plant small RNAs. Biogenesis pathways of micro RNAs (miRNAs), *trans*-acting siRNAs (tasiRNAs), natural-antisense-transcript derived siRNAs (natsiRNAs), and *cis*-acting siRNAs (casiRNAs) are composed of different RNAi component genes. Their precursor dsRNAs are either synthesized by RDRs or formed by internal palindrome sequences and are further excised by DCL1–4 in distinguishable modes[7]. RISC, RNA-induced silencing complex; RdDM, RNA-directed DNA methylation. (B) An example locus of casiRNA-deriving Pri-TU inside a transposable element (AT1TE00835). Each bar represents a smRNA-Seq read differentially colored by the 5′ terminal nucleotide (details of *ab initio* identification of Pri-TUs are described in Supplementary Methods).

Please cite this article as: X. Wang, et al., Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites, Genomics (2011), doi:10.1016/j.ygeno.2011.01.006

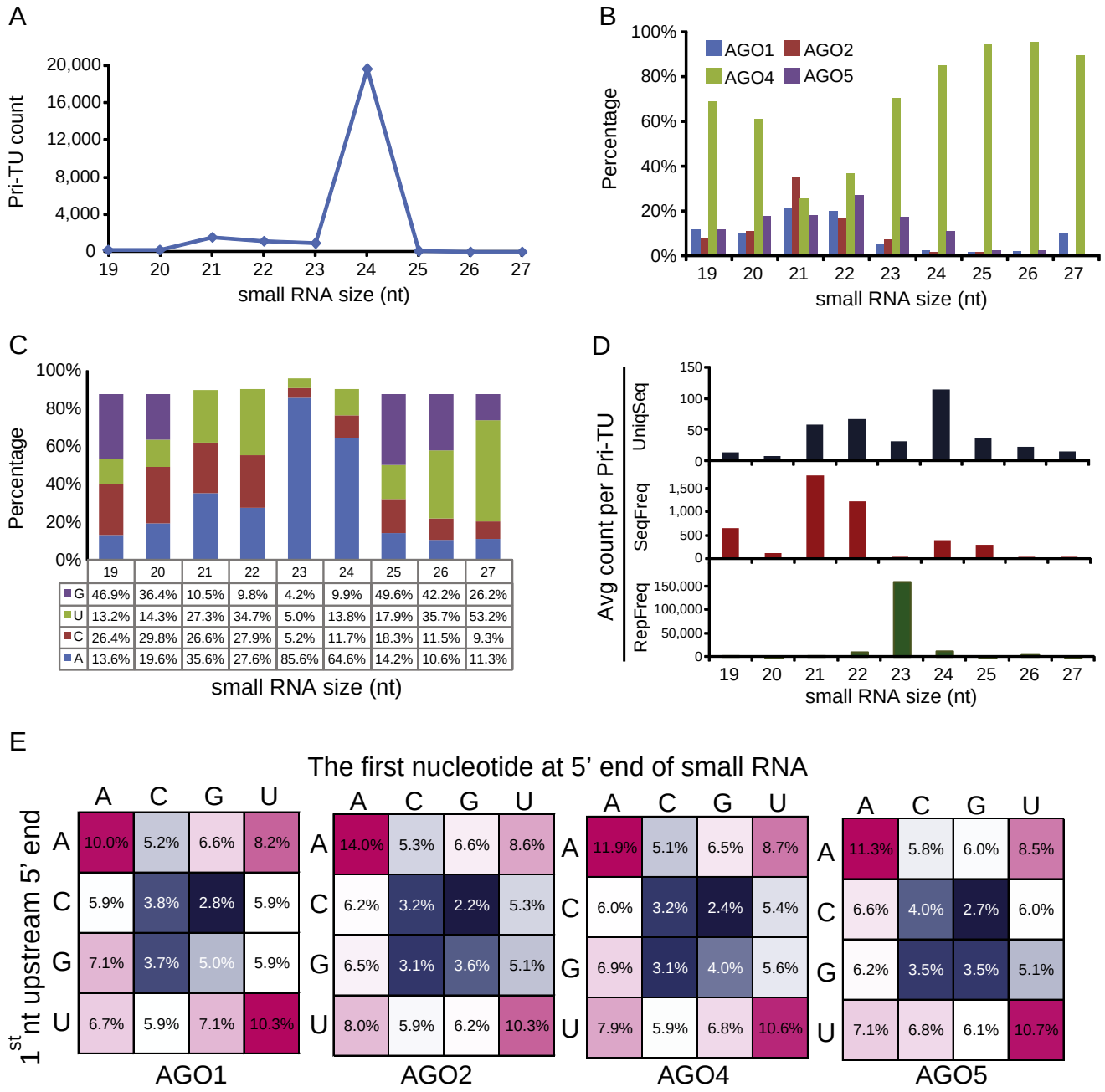

Fig. 2. Sequence characteristics of the smRNAs in the Pri-TUs. (A) Pri-TUs predominantly produce 24-nt small RNAs. (B) Preferential association of small RNAs in different AGO complexes. (C) Preferential type of 5′ terminal nucleotide in different size classes of small RNAs. (D) SeqFreq, RepFreq and UniqFreq (definitions were described in the main text) from the small RNA reads within a given Pri-TU to represent its expression abundance, repetitiveness and excision mode, respectively. (E) The frequencies of di-nucleotide cutting sites (the first and the first upstream nucleotide at a small RNA's 5′ terminus) in different AGOs.

candidates for the further prediction of new miRNA genes (Fig. 3C). Initially, 632 Pri-TUs were predicted as harboring putative miRNAs, which encompassed 113 of 118 (95.7%) known *Arabidopsis* miRNA genes. We further examined the 519 candidate Pri-TUs not containing previously annotated miRNA genes. Interestingly, we found the tasiRNA genes were among the candidates, as they were also precisely processed from a tasiRNA precursor in a phased mode, and many copies of reads were sequenced (Supplementary Fig. 8). This feature made the PCA unable to distinguish them from the miRNAs. Other false positive candidate Pri-TUs were actually associated with tRNA, rRNAs, snRNAs and snoRNAs etc. Those RNAs possess internal stem–loop structure, and some of the regions have higher frequency to be degraded. Although the non-canonical miRNAs have been repeatedly reported that they were produced from tRNA, rRNA or snoRNAs, however, from the distribution of the size and 5′ terminal nucleotide, they did not resemble to canonical miRNAs or TE-derived siRNAs but were likely to be the functionless degradation products (Supplementary Fig. 8). Filtering these Pri-TUs further narrowed down the candidates to only 36 Pri-TUs, which are located unambiguously in the intergenic regions. The smRNA-Seq reads from the 36 Pri-TUs possess miRNA-like features, as 64% of them are associated with AGO1 (Fig. 3D), and the predominant class is 21 nt with 5′ terminal U (Fig. 3E). Based on the fact that canonical miRNAs are processed from the shRNA precursors [34], we extracted the putative precursor

Please cite this article as: X. Wang, et al., Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites, Genomics (2011), doi:10.1016/j.ygeno.2011.01.006

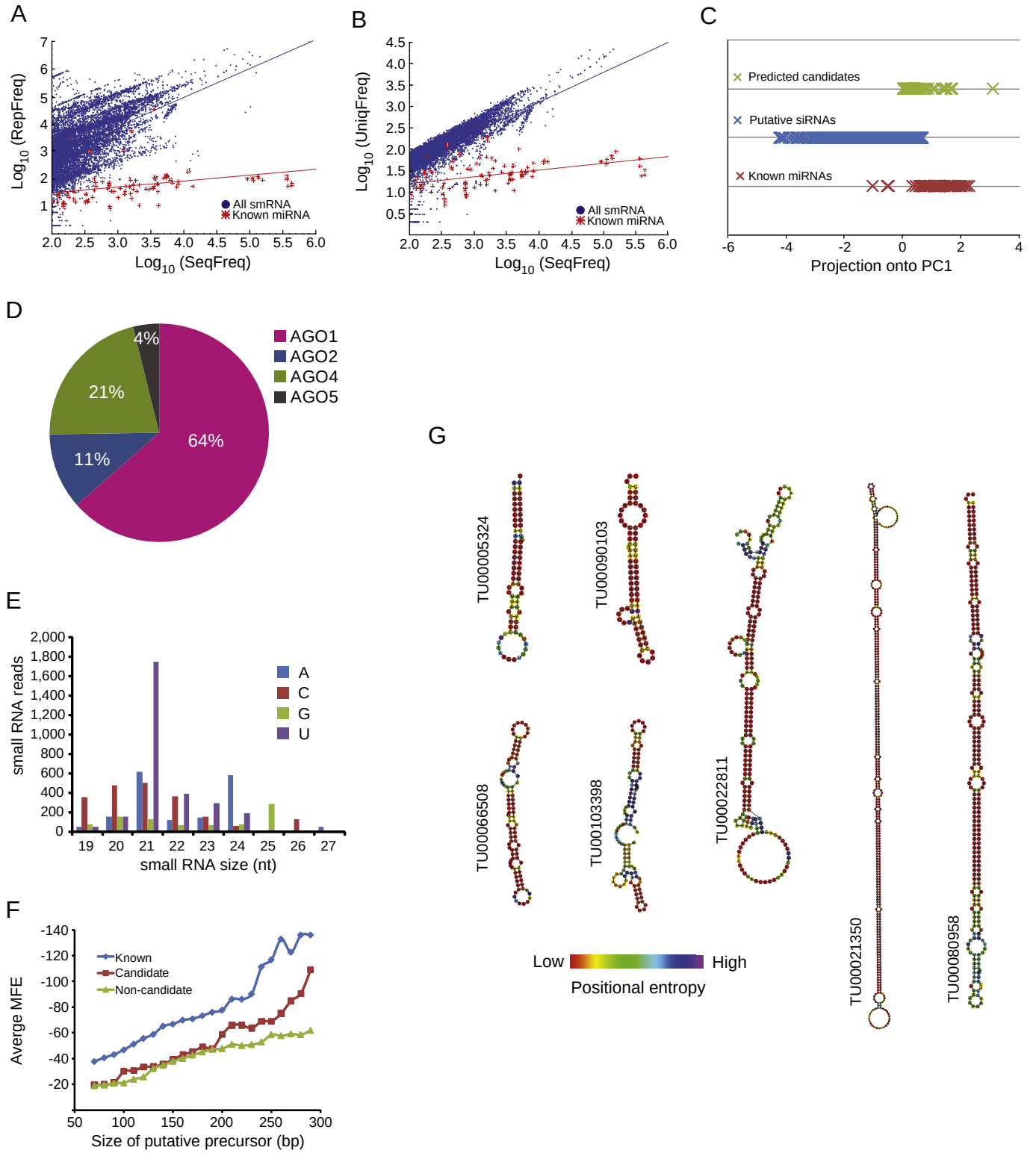

Fig. 3. Computational selection of candidate Pri-TUs for new *MIR* gene prediction. (A) and (B) The Pri-TUs encoding known miRNAs tend to have higher SeqFreq, but lower RepFreq and UniqFreq. (C) Three groups of Pri-TUs projected onto the first principal component (PC1). (D) and (E) Most of the small RNA reads from the 36 candidate Pri-TUs are preferentially associated with AGO1, 21 nt in length and initiated with a 5′ U. (F) The relationship between precursor length and average minimum free energy (MFE) in groups of known miRNAs, candidate and non-candidate Pri-TUs. (G) Examples of the secondary structures of selected putative *MIR* genes whose RNA precursors were capable of forming either short hairpin or long hairpin stem–loop structure.

sequences (minimum 70 nt) of 36 candidates for 2nd structure prediction by RNAfold [35]. Since the minimum free energy (MFE) anti-correlates with the lengths of precursor RNAs, we compared the MFE calculated from the 118 known miRNA precursors, the 36 candidate Pri-TU, and randomly selected ~100 non-candidate Pri-TU sequences (size ≥ 70 and < 300 nt). While the MFE of known miRNAs was significantly lower than non-candidate group, the 36 candidates were between the two groups when the precursor size was above 200 nt (Fig. 3F). This pattern suggests that some previously unannotated miRNAs might have longer precursors than canonical pre-

Please cite this article as: X. Wang, et al., Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites, Genomics (2011), doi:10.1016/j.ygeno.2011.01.006

miRNA hairpins. A selection of Pri-TUs harboring putative novel miRNAs and their second structures are shown in Fig. 3G.

3.3. Long hairpin dsRNAs formed by gene and TE pairs produce different types of small RNAs

The existence of long hairpin dsRNAs (lh-dsRNAs) has been reported in both plants and animals [36,37], and its functional significance is emphasized as an alternative pathway of smRNAs biogenesis which is independent of the activity of RNA-DEPEDNET RNA POLYMERASE (*RDR*) [38,39]. To characterize the function and components of RDR-independent pathways, we investigated the patterns of smRNA production in different *Arabidopsis* developmental stages and mutation backgrounds. Using *einverted* [40], we first *de novo* predicted 2674 genomic loci with potentials of forming stem–loop structures, and then focused on 15 high scoring ones selected by the stringent criteria: (a) stem length ≥ 500 nt, (b) stem identity ≥ 90%, and (c) loop length ≤ 300 nt. Surprisingly, the lh-dsRNAs loci are formed by various genomic elements such as TEs, centromeric repeats, protein-coding genes or 5S rRNAs, which produced smRNAs of different sizes (Supplementary Fig. 9). Our attention was attracted to two lh-dsRNAs formed by distinct sources, a pair of TEs (AT2TE23505/AT2TE23510) and a pair of protein-coding genes (AT3G44570/AT3G44580) (Fig. 4A and B). The TE pair located in chromosome 2 centromere predominantly produced 24-nt siRNAs in association with AGO4, and exhibited strong strand-bias (Fig. 4C). In addition, smRNA production from this TE pair was partially independent of RDR2, and very sensitive to the triple mutation of *dcl2/3/4*, but not to the mutation of *dcl1*. In contrast, the gene pair showed a scenario of miRNA-like biogenesis: first, the prevalent size classes of smRNAs were 21 and 22 nt yielded from both strands of the lh-dsRNA, mostly in association with AGO1; secondly, production of smRNAs was independent of DCL2/3/4 and RDR2, but extremely sensitive to *dcl1* mutant. What is more, these two distinct lh-dsRNAs demonstrated different patterns in DNA methylation status [41], tissue-specific productions, and responses to *met1*, *ddc*, *rdd* mutation backgrounds (Fig. 4C and D).

Our analysis suggests that lh-dsRNA formed by the TE pair entered the siRNA biogenesis pathway, whereas the one by gene pair entered a pathway resembling miRNA biogenesis. We hypothesize that the fundamental differences between TE-formed and gene-formed lh-dsRNAs may initially arise from their transcription by different plant polymerases, Pol IV/V and Pol II, respectively. As a matter of fact, both AT3G44570 and AT3G44580 are annotated as "hypothetical protein" without known functions, and no expression signals were detected in any developmental stage (Supplementary Figs. 10A and B) [42] More interestingly, a detailed analysis of AT3G44570 identified a TE insertion domain in it, which interrupted the ORFs and was the probable cause of the decay of this gene (Supplementary Fig. 10C). A recent model proposed that the inverted gene duplication may be the evolutionary origin of a modern *MIR* gene, and its fate into DCL3/AGO4 or DCL1/AGO1 pathways was adaptively selected by the bugles in the dsRNA stem–loop acquired from mutations [1,43]. We hypothesized that the lh-dsRNA formed by decayed AT3G44570/AT3G44580 pair is a vivid prototype of an evolving miRNA gene.

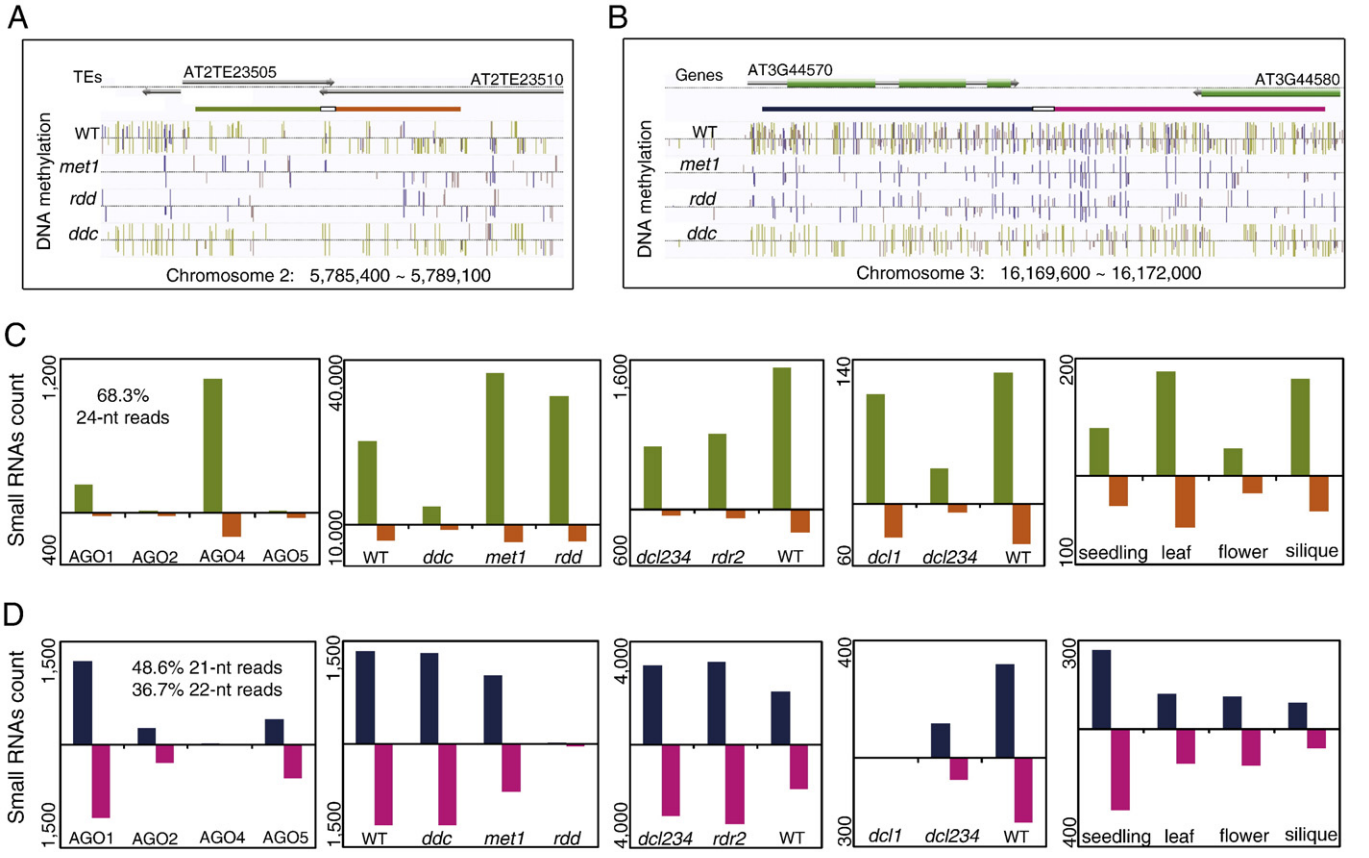

Fig. 4. Distinct patterns of small RNA biogenesis from two long hairpin dsRNAs formed by inverted duplications of TEs and protein-coding genes. (A) Two centromeric TEs, AT2TE23505 and AT2TE23510, form a long hairpin dsRNA with hypomethylation in wild type. Green and orange bars represent the stem regions, and the white part is the loop region. (B) Two non-TE genes, AT3G44570 and AT3G44580, form a long hairpin dsRNA with hypermethylation in wild type. Blue and pink bars represent the stem regions, and the white part is the loop region. (C) Small RNA production of the studied TE pair in different developmental stages, mutation backgrounds and in association with AGO complexes. (D) Small RNA production of the studied gene pair in different developmental stages, mutation backgrounds and in association with AGO complexes.

Please cite this article as: X. Wang, et al., Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites, Genomics (2011), doi:10.1016/j.ygeno.2011.01.006

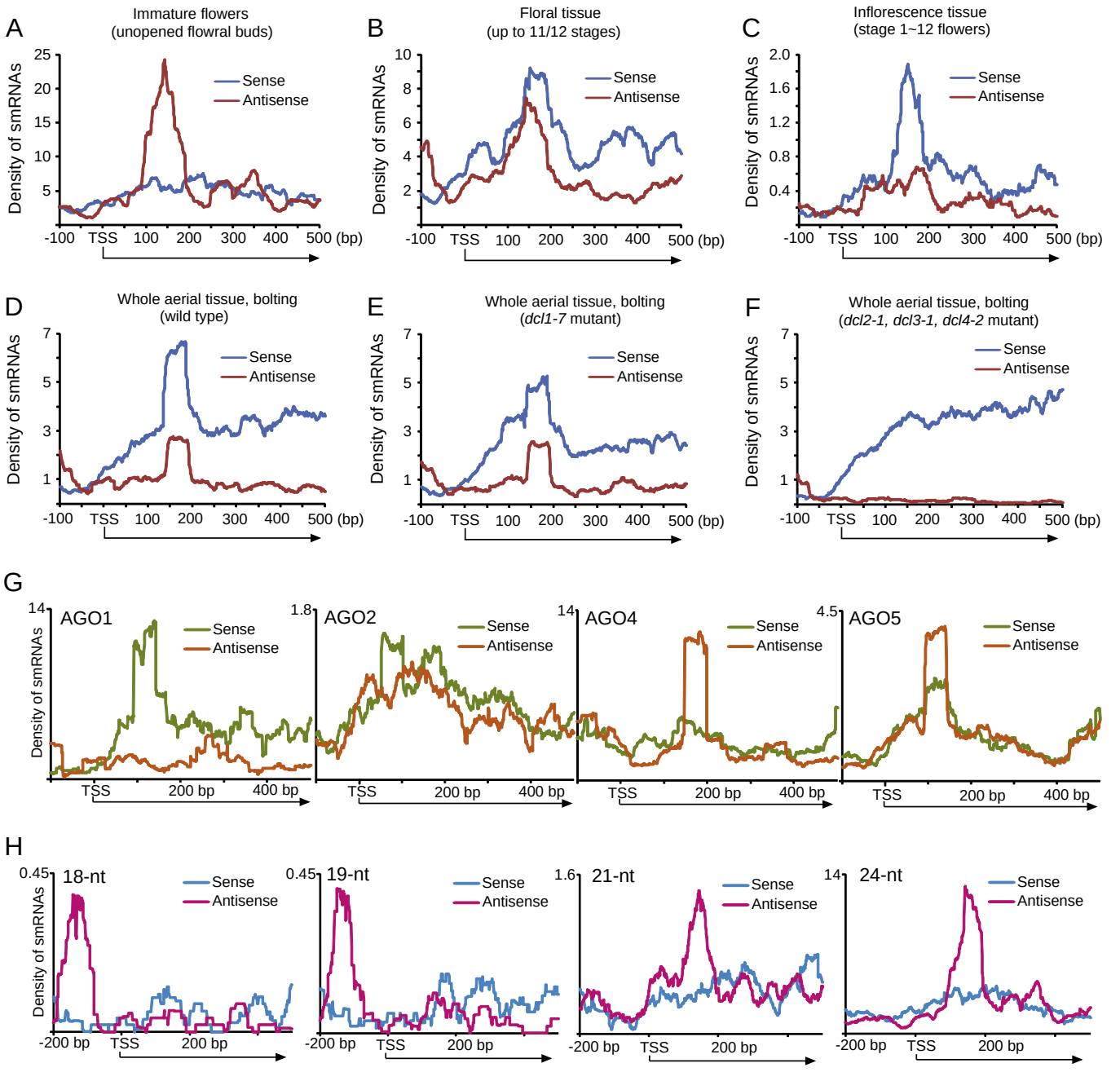

Fig. 5. Sense and antisense TSSa-RNAs are produced from siRNA biogenesis pathways. (A), (B) and (C) Developmental change in proportion of sense and antisense TSSa-RNAs peaked from 100 to 200 bp downstream TSSs (smRNA-Seq data in wild type from A: GSM277608, B: GSM280228, and C: GSM154336). (D), (E) and (F) Biogenesis of TSSa-RNAs is sensitive to *dcl2/3/4* triple mutant, but not *dcl1* mutant (D: GSM366868, E: GSM366869, and F: GSM366870). (G) Sense and antisense TSSa-RNAs are differentially associated with AGO1 and AGO4, respectively (GSE10036). See Supplementary Table 1 for detailed description of the plant materials. (H) The antisense promoter-associated small RNAs (PASRs) are located from 100- to 200-bp upstream TSSs, and shorter than antisense TSSa-RNAs (GSM277608).

3.4. Biogenesis pathways of small RNAs involved in transcription initiation

Short RNAs associated with transcription start sites (TSSa-RNAs) have been recently reported in animals, which were found positively correlated with gene transcription initiation (or named tiRNAs) [44,45]. In addition, promoter-associated short RNAs (PASRs), especially for those occurring on the antisense strand of the promoters, may establish and maintain the long-term transcriptional silencing of the nearby genes in human cells [46,47]. We were curious about the existence of these two types of small RNAs in plants, even though it has been reported absent in plants [45]. We explored their potential function and pathways in different *Arabidopsis* tissues and mutation backgrounds. To exclude the potential influences from pseudogenes, we focused this analysis on the ~17,000 non-TE genes with definitive functional descriptions. We first mapped the three sets of smRNA-Seq reads to the selected genes within the range between upstream 100 bp and downstream 500 bp from the TSSs, and calculated the average density of smRNA reads in a sliding 50 bp window. To reduce the potential bias caused by reads derived from repetitive regions, we then averaged the read density in the 50 bp window by dividing the number of genomic locations the reads were mapped to.

Surprisingly, the three studied datasets demonstrated distinct patterns of smRNA abundance within the downstream 100 to 200 bp region from the aligned TSSs (Fig. 5). While the smRNAs in the first

and third dataset exhibited solitary antisense and sense peaks (Fig. 5A and C), respectively, the second dataset showed both peaks (Fig. 5B). The cause of the discrepancy might be that three smRNA-Seq were conducted in different flower developmental stages. These observations suggest that two types of TSSa-RNAs might play regulatory roles in transcription initiation, namely the sense (+) TSSa-RNAs and the antisense (−) TSSa-RNAs.

To identify the genes regulating the biogenesis of TSSa(+/−)-RNAs, we compared the TSSa-RNAs abundances on the same 17,000 genes in wild type, *dcl1* mutant and *dcl2/3/4* triple mutant plants (Fig. 5). We observed the coexistence of sense and antisense peaks of TSSa-RNAs in wild type (Fig. 5D), slight decrease of the sense peak but unchanged antisense peak in *dcl1* mutant (Fig. 5E), and the complete abolishment of TSSa-RNA peaks in *dcl2/3/4* triple mutant (Fig. 5F). This result indicates that the siRNA biogenesis pathways instead of the miRNA ones are responsible for producing the TSSa-RNAs. It was further evidenced by the examination of the association of TSSa(+/−) a-RNAs with different AGO complexes (Fig. 5G). While most of TSSa (+)-RNAs were found in AGO1, the TSSa(−)-RNAs were preferentially in AGO4. AGO5 seems capable of associating with both sense and antisense TSSa-RNA, but TSSa-RNAs are depleted of AGO2. At last, we detected the antisense promoter-associated small RNAs by expanding the upstream region to 200 bp from TSSs. Interestingly, antisense PASRs were in low abundance, and shorter (18–20 nt) than antisense TSSa-RNAs (21–24 nt) located downstream the TSSs (Fig. 5H and Supplementary Fig. 11).

Taft et al. reported the absence of promoter- and TSS-associated RNAs in *Arabidopsis* [45]. However, we indeed detected them from multiple datasets, despite their low abundance. We reasoned that Taft et al. [45] probably used the full set of ~30,000 genes with considerable proportion of pseudogenes or TEs, and did not normalize the bias from TE-derived repetitive siRNAs, which might veil the real patterns on *bona fide* genes. TSSa-RNAs and PASRs have been thought to originate from the frequent divergent transcription of pre-engaged RNA Pol II in animals [48,49]. In plants, Pol IV and Pol V produce the non-coding RNAs that trigger the epigenetic silencing machinery on overlapping or neighboring genes [49]. We provided the supporting evidences for these postulations and demonstrated that the RNAi pathway related genes might be involved in the function and biogenesis of TSSa-RNAs during the epigenetic regulation of transcription initiation.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.01.006.

Conflict of interest statement

None declared.

Acknowledgments

We thank Dr. Yijun Qi of the National Institute of Biological Sciences at Beijing for generously providing the processed AGO1/2/4/5 associated smRNA-Seq data. We are grateful to Scott Taing for proofreading the manuscript. Dr. Xiangfeng Wang is a research fellow supported by Sloan Research Fellowship and NIH grant HG004069.

References

[1] O. Voinnet, Origin, biogenesis, and activity of plant microRNAs, Cell 136 (4) (2009) 669–687.
[2] H. Vaucheret, M. Fagard, Transcriptional gene silencing in plants: targets, inducers and regulators, Trends Genet. 17 (1) (2001) 29–35.
[3] F. Vazquez, H. Vaucheret, R. Rajagopalan, C. Lepers, V. Gasciolli, A.C. Mallory, J.L. Hilbert, D.P. Bartel, P. Crété, Endogenous trans-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs, Mol. Cell 16 (1) (2004) 69–79.
[4] Z. Xie, L.K. Johansen, A.M. Gustafson, K.D. Kasschau, A.D. Lellis, D. Zilberman, S.E. Jacobsen, J.C. Carrington, Genetic and functional diversification of small RNA pathways in plants, PLoS Biol. 2 (5) (2004) E104.
[5] D. Zilberman, X. Cao, S.E. Jacobsen, ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation, Science 299 (2003) 716–719.
[6] A. Kloc, M. Zaratiegui, E. Nora, R. Martienssen, RNA interference guides histone modification during the S phase of chromosomal replication, Curr. Biol. 18 (7) (2008) 490–495.
[7] D.J. Obbard, D.J. Finnegan, RNA interference: endogenous siRNAs derived from transposable elements, Curr. Biol. 18 (13) (2008) R561–R563.
[8] S. Chan, D. Zilberman, Z. Xie, L.K. Johansen, J.C. Carrington, S.E. Jacobsen, RNA silencing genes control de novo DNA methylation, Science 303 (2004) 1336.
[9] M. Ghildiyal, P.D. Zamore, Small silencing RNAs: an expanding universe, Nat. Rev. Genet. 10 (2) (2009) 94–108.
[10] I.R. Henderson, X. Zhang, C. Lu, L. Johnson, B.C. Meyers, P.J. Green, S.E. Jacobsen, Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning, Nat. Genet. 38 (2006) 721–725.
[11] Y. Onodera, J.R. Haag, T. Ream, P.C. Nunes, O. Pontes, C.S. Pikaard, Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation, Cell 120 (5) (2005) 613–622.
[12] C.S. Pikaard, J.R. Haag, T. Ream, A.T. Wierzbicki, Roles of RNA polymerase IV in gene silencing, Trends Plant Sci. 13 (7) (2008) 390–397.
[13] H. Vaucheret, Plant ARGONAUTES, Trends Plant Sci. 13 (7) (2008) 350–358.
[14] B.J. Reinhart, E.G. Weinstein, M.W. Rhoades, B. Bartel, D.P. Bartel, MicroRNAs in plants, Genes Dev. 16 (13) (2002) 1616–1626.
[15] E. Allen, Z. Xie, A.M. Gustafson, J.C. Carrington, microRNA-directed phasing during trans-acting siRNA biogenesis in plants, Cell 121 (2) (2005) 207–221.
[16] O. Borsani, J. Zhu, P.E. Verslues, R. Sunkar, J.K. Zhu, Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in *Arabidopsis*, Cell 123 (7) (2009) 1279–1291.
[17] Y. Onodera, J.R. Haag, T. Ream, P.C. Nunes, O. Pontes, et al., Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation, Cell 120 (2005) 613–622.
[18] A.T. Wierzbicki, J.R. Haag, C.S. Pikaard, Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes, Cell 135 (4) (2008) 635–648.
[19] Ian R. Henderson, Steven E. Jacobsen, Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading, Genes Dev. 22 (2008) 1597–1606.
[20] W.C. Wang, F.M. Lin, W.C. Chang, K.Y. Lin, H.D. Huang, N.S. Lin, miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression, BMC Bioinform. 10 (2009) 328.
[21] P.J. Huang, Y.C. Liu, C.C. Lee, W.C. Lin, R.R. Gan, P.C. Lyu, P. Tang, DSAP: deep-sequencing small RNA analysis pipeline, Nucleic Acids Res. 1 (Jul 2010) 38.
[22] M.R. Friedländer, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, N. Rajewsky, Discovering microRNAs from deep sequencing data using miRDeep, Nat. Biotechnol. 26 (4) (2008) 407–415.
[23] J.H. Yang, P. Shao, H. Zhou, Y.Q. Chen, L.H. Qu, deepBase: a database for deeply annotating and mining deep sequencing data, Nucleic Acids Res. 38 (2010) D123–D130, (Database issue).
[24] E. Zhu, F. Zhao, G. Xu, H. Hou, L. Zhou, X. Li, Z. Sun, J. Wu, mirTools: microRNA profiling and discovery based on high-throughput sequencing, Nucleic Acids Res. 38 (2010) W392–W397, (Web Server issue).
[25] M. Hackenberg, M. Sturm, D. Langenberger, J.M. Falcón-Pérez, A.M. Aransay, miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments, Nucleic Acids Res. 37 (2009) W68–W76, (Web Server issue).
[26] S. Mi, T. Cai, Y. Hu, Y. Chen, E. Hodges, F. Ni, L. Wu, S. Li, H. Zhou, C. Long, S. Chen, G.J. Hannon, Y. Qi, Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5′ terminal nucleotide, Cell 133 (1) (2008) 116–127.
[27] R. Lister, R.C. O'Malley, J. Tonti-Filippini, B.D. Gregory, C.C. Berry, A.H. Millar, J.R. Ecker, Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*, Cell 133 (3) (2008) 523–536.
[28] M.A. German, M. Pillay, D.H. Jeong, A. Hetawal, S. Luo, P. Janardhanan, V. Kannan, L.A. Rymarquis, K. Nobuta, R. German, E. De Paoli, C. Lu, G. Schroth, B.C. Meyers, P.J. Green, Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends, Nat. Biotechnol. 26 (8) (2008) 941–946.
[29] K.D. Kasschau, N. Fahlgren, E.J. Chapman, C.M. Sullivan, J.S. Cumbie, S.A. Givan, J.C. Carrington, Genome-wide profiling and analysis of *Arabidopsis* siRNAs, PLoS Biol. 5 (3) (2007) e57.
[30] N. Fahlgren, C.M. Sullivan, K.D. Kasschau, E.J. Chapman, J.S. Cumbie, T.A. Montgomery, S.D. Gilbert, M. Dasenko, T.W. Backman, S.A. Givan, J.C. Carrington, Computational and analytical framework for small RNA profiling by high-throughput sequencing, RNA 15 (5) (2009) 992–1002.
[31] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (3) (2009) R25.
[32] M. Ringnér, What is principal component analysis? Nat. Biotechnol. 26 (3) (2008) 303–304.
[33] P. Brodersen, O. Voinnet, The diversity of RNA silencing pathways in plants, Trends Genet. 22 (5) (2006) 268–280.
[34] B.C. Meyers, M.J. Axtell, B. Bartel, D.P. Bartel, D. Baulcombe, J.L. Bowman, X. Cao, J.C. Carrington, X. Chen, P.J. Green, S. Griffiths-Jones, S.E. Jacobsen, A.C. Mallory, R.A. Martienssen, R.S. Poethig, Y. Qi, H. Vaucheret, O. Voinnet, Y. Watanabe, D. Weigel, J.K. Zhu, Criteria for annotation of plant MicroRNAs, Plant Cell 20 (12) (2008) 3186–3190.
[35] I.L. Hofacker, Vienna RNA secondary structure server, Nucleic Acids Res. 31 (13) (2003) 3429–3431.
[36] X. Zhang, I.R. Henderson, C. Lu, P.J. Green, S.E. Jacobsen, Role of RNA polymerase IV in plant small RNA metabolism, Proc. Natl Acad. Sci. USA 104 (11) (2007) 4536–4541.

Please cite this article as: X. Wang, et al., Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites, Genomics (2011), doi:10.1016/j.ygeno.2011.01.006

[37] O.H. Tam, A.A. Aravin, P. Stein, A. Girard, E.P. Murchison, S. Cheloufi, E. Hodges, M. Anger, R. Sachidanandam, R.M. Schultz, G.J. Hannon, Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes, Nature 453 (7194) (2008) 534–538.
[38] K. Nobuta, C. Lu, R. Shrivastava, M. Pillay, E. De Paoli, M. Accerbi, M. Arteaga-Vazquez, L. Sidorenko, D.H. Jeong, Y. Yen, P.J. Green, V.L. Chandler, B.C. Meyers, Distinct size distribution of endogeneous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. Proc. Natl Acad. Sci. USA 105 (39) (2009) 14958–14963.
[39] X. Wang, A.A. Elling, X. Li, N. Li, Z. Peng, G. He, H. Sun, Y. Qi, X.S. Liu, X.W. Deng, Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize, Plant Cell 21 (4) (2009) 1053–1069.
[40] http://emboss.bioinformatics.nl/cgi-bin/emboss/einverted/.
[41] http://neomorph.salk.edu/epigenome/epigenome.html.
[42] M. Schmid, T.S. Davison, S.R. Henz, U.J. Pape, M. Demar, M. Vingron, B. Schölkopf, D. Weigel, J.U. Lohmann, A gene expression map of *Arabidopsis thaliana* development, Nat. Genet. 37 (5) (2005) 501–506.
[43] E. Allen, Z. Xie, A.M. Gustafson, G.H. Sung, J.W. Spatafora, J.C. Carrington, Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*, Nat. Genet. 36 (2004) 1282–1290.
[44] R.J. Taft, C. Simons, S. Nahkuri, H. Oey, D.J. Korbie, T.R. Mercer, J. Holst, W. Ritchie, J.J. Wong, J.E. Rasko, D.S. Rokhsar, B.M. Degnan, J.S. Mattick, Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans, Nat. Struct. Mol. Biol. 17 (8) (2010) 1030–1034.
[45] R.J. Taft, E.A. Glazov, N. Cloonan, C. Simons, S. Stephen, G.J. Faulkner, T. Lassmann, A.R. Forrest, S.M. Grimmond, K. Schroder, K. Irvine, T. Arakawa, M. Nakamura, A. Kubosaki, K. Hayashida, C. Kawazu, M. Murata, H. Nishiyori, S. Fukuda, J. Kawai, C.O. Daub, D.A. Hume, H. Suzuki, V. Orlando, P. Carninci, Y. Hayashizaki, J.S. Mattick, Tiny RNAs associated with transcription start sites in animals, Nat. Genet. 41 (5) (2009) 572–578.
[46] J. Han, D. Kim, K.V. Morris, Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells, Proc. Natl Acad. Sci. USA 104 (30) (2007) 12422–12427.
[47] P.G. Hawkins, S. Santoso, C. Adams, V. Anest, K.V. Morris, Promoter targeted small RNAs induce long-term transcriptional gene silencing in human cells, Nucleic Acids Res. 37 (9) (2009) 2984–2995.
[48] P. Kapranov, J. Cheng, S. Dike, D.A. Nix, R. Duttagupta, A.T. Willingham, P.F. Stadler, J. Hertel, J. Hackermüller, I.L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammana, T.R. Gingeras, RNA maps reveal new RNA classes and a possible function for pervasive transcription, Science 316 (5830) (2007) 1484–1488.
[49] A.C. Seila, J.M. Calabrese, S.S. Levine, G.W. Yeo, P.B. Rahl, R.A. Flynn, R.A. Young, P.A. Sharp, Divergent transcription from active promoters, Science 322 (5909) (2008) 1849–1851.

Please cite this article as: X. Wang, et al., Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites, Genomics (2011), doi:10.1016/j.ygeno.2011.01.006